DEVELOPING AN ITEM BANK FOR HOMOGENEOUS SECOND ORDER DIFFERENTIAL EQUATIONSBY CALIBRATED ITEMS

Elahe Aminifar, PhD
Shahid Rajaee Teacher Training University, Iran
Mohammad Alipour, MSc Student
Shahid Rajaee Teacher Training University, Iran

Abstract

Item bank is one of the main components of adaptive tests. In this research, a test was made in order to design and calibrate items for Homogeneous Second Order Differential Equations. The items were designed according to the goal-content's table of the subject and the Bloom's taxonomy learning domain. Validity and reliability of these items was confirmed by academic staff who have taught the course for years. For calibrating items, 13 levels of ability were considered. By using Monte Carlo simulation, 32500 simulated examinees (2500 simulated examinees for each ability level) participated in the exam. Calibrating items were done by difficulty and discrimination parameters using item response theory and priory method. The results showed that chi-square indices of parameters is less than the standard chi-square indices, and therefore the estimated parameters are acceptable. These items can be used in adaptive tests in order to estimate examinee's ability level in this subject.

Keywords: Item calibration, Difficulty parameter, Discrimination parameter, Ability level

Introduction

Currently universities usually use conventional fixed length tests to measure students' ability. This means, they ask a set of items from all students and then compare their scores. The main disadvantage of this method is when a student's ability is far from the difficulty of the test, thushis score is measured inaccurately (Weiss, 2011). In adaptive tests, by asking items with proper difficulty from each student, the standard error of measurement will decrease. One of the main components of adaptive tests is

the item bank, that consists of calibrated items from which the items of test will be selected from it. Obviously, without designing a proper item bank, administering an adaptive test will become impossible.

Adaptive Test

An adaptive test has six main components:

- Theory of measurement: A network of hypotheses and deductions associated with the construct we are attempting to measure (Sympson, 1970). Two main theories in adaptive tests are Classical Test Theory (CTT) and Item Response Theory (IRT) (Weiss, 2011).
- Item bank: A set of items which their psychometric properties like difficulty, discrimination and guessing parameters are calibrated (Thompson & Weiss, 2011).
- Starting point (Initial ability): Before the test begins, based on the prior information about student, a number will be assigned to him as his initial ability. The first item of the student depends on his initial ability (Weiss, 2004).
- Item selection algorithm: This algorithm selects the next item of test from the item bank (van der Linden, 2005).
- Scoring algorithm: After answering an item, the scoring algorithm estimates the student's new estimated ability according to his answer (Weiss, 2011).
- Termination criteria: The conditions which their satisfaction will terminate the testing process(Babcock& Weiss, 2009).

In the present study, we design an item bank for homogeneous second order differential equations that can be implemented in an adaptive test.

Item Response Theory (IRT)

The main property of Item Response Theory is estimating the chance that a person with θ (ability) level answers an item with b (difficulty) parameter correctly(Thompson & Weiss, 2011). Obviously, in two-parameter model of IRT, the chance that a person with θ (ability) level answers an item with b (difficulty) and a (discrimination) parameters will be correctly estimated (Baker, 2001). This chance can be calculated by:

$$P(\theta) = \frac{1}{1+e^{a(b-\theta)}} \qquad (1)$$

This function is also called as item characteristic curve.

Priory method

Assume M students answered N items of the test. In this method, the initial abilities of students were calculated from their final score of the test.

For comfort, we consider *J* levels of ability (ranging from -3 to 3) and distribute all students into these levels. In level *j*, there will be m_j students (*j* = 1, 2… *J*).

For estimating an item's parameters, the proportion of students in *j* level that have answered the item correctly, denoted by $P_O(\theta_j)$, is considered as an estimate of $P(\theta_j)$. The same process will be repeated for all*j* = 1, 2… *J*.

Now initial values for the item parameters, such as *b* = 0.0, *a* = 1.0, are established as *a priori*. Then, using these parameters, the value of $P(\theta_j)$ is computed at each ability level. The agreement of the observed value of $P_O(\theta_j)$ and computed value $P(\theta_j)$ is determined across all ability groups. If there was significant difference between $P_O(\theta_j)$ and $P(\theta_j)$, then the *b* and *a* parameters will change and the same process will repeat. This process of adjusting the parameters is continued until the adjustments get so small that little improvement in the agreement is possible. One can select the *b* and *a* parameters which leads to the least sum square of difference between $P_O(\theta_j)$ and $P(\theta_j)$.At this point, the estimation procedure is terminated and the current values of *b* and *a* are estimations of the item parameter.

An important consideration within item response theory is whether a particular item characteristic curve model fits the item response data for an item. The agreement of the observed proportions of correct response and those yielded by the fitted item characteristic curve for an item is measured by the chi-square goodness-of-fit index. This index is defined as follows:

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{[P_o(\theta_j) - P(\theta_j)]^2}{P(\theta_j)Q(\theta_j)} \qquad (2)$$

Where *J* is the number of ability groups,$\theta_j$is the ability level of group *j*, $m_j$is the number of students having ability θ_j, $P_O(\theta_j)$is the observed proportion of correct response for group *j*, $P(\theta_j)$ is the probability of correct response for group *j* computed from the item characteristic curve model using the item parameter estimates and $Q(\theta_j)$ is equal to 1- $P(\theta_j)$.

If the value of the obtained index is greater than a criterion value, the item characteristic curve specified by the values of the item parameter estimates will not fit the data (Baker, 2001).

Monte Carlo simulation

Monte Carlo simulation is based on the fact that IRT provides an estimate of the exact probability of a correct response to an item for a given value of θ (Thompson & Weiss, 2011). For example, suppose that an item has estimated to have *b* = 0 difficulty and *a* = 1 discrimination parameters. The first simulated examinee with arbitrary ability (usually -3) will be considered. The probability of the correct response is:

$$P(\theta) = \frac{1}{1 + e^{1(0-(-3))}} = 0.05$$

Now a random number is generated from a uniform distribution with a range of 0 to 1. If the number is 0.05 or less, the first simulated examinee is supposed to give a correct answer. Otherwise, the answer is incorrect.

This process repeats for simulated examinees with various ability levels and their responses will be gathered.

Methodology

The present study offers calibrated items for developing the item bank of homogeneous second order differential equations. Initially, 61 items were designed by researchers according to the goal-content'stable of the subject and the Bloom's taxonomy learning domain. Validity and reliability of these items were confirmed by academic staff who have taught the course several times. For calibrating items, 13 levels of ability were considered. By using Monte Carlo simulation, 32500 simulated examinees (2500 simulated examinees for each ability level) participated in the exam. The initial values for items' difficulty and discrimination parameters were selected by academic staff who have taught the course several times. After generating simulated examinees' answers, items were calibrated for difficulty and discrimination parameters using item response theory and priory method. Finally, the agreement of the observed proportions of correct response and those yielded by the fitted item characteristic curve for the item was measured by the chi-square goodness-of-fit index.

A practical example of calibrating an item

In the present study, M = 32500 examinees answered N = 61 items. The ability scale (from -3 to 3) has been divided into 12 equal pieces which leads to 13 ability levels. In each ability level, there were 2500 simulated examinees. These ability levels and their ability values (θ_j) are shown in Table 1.

Table 1. The value of ability in each level

Ability level	1	2	3	4	5	6	7	8	9	10	11	12	13
Ability Value	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3

For example, the process of generating simulated examinees' answers and calibrating the item 16 will be described step-by-step:

For generating simulated examinees' answers, the initial values for difficulty (*b*) and discrimination parameters (*a*)were asked from two

academic staff who have taught the course several times. One of them suggests b = 1, a = 1 and the other one suggests b = 0.5, a = 0.7. So, the mean of these values, b = 0.75, a = 0.85, were accepted as the initial values for difficulty and discrimination parameters. These values will be used in the Monte Carlo simulation as follows. The first simulated examinee with ability -3 will be considered. The probability of the correct response is:

$$P(\theta) = \frac{1}{1 + e^{0.85(0.75-(-3))}} = 0.04$$

Now a random number is generated from a uniform distribution with a range of 0 to 1.The generated number is 0.43. Since 0.43is greater than 0.04, the generated answer will be supposed incorrect.

For generating the next 2499 answers, the ability level (θ) of simulated examinees will establish on -3, and by generating random numbers, they will be compared to 0.04. After generating the first 2500 answers, the ability level (θ) of simulated examinees will establish on -2.5, and after calculating the probability of correct answer, *P(θ)*, the next 2500 random numbers will be compared to it. This procedure will repeat for all ability levels in order to generate all 32500 answers for item 16.

After generating the simulated examinees' answers, the examinees that answered the item 16 correctly will be classified by their ability levels. In this study, 11563 examinees have answered correctly to item 16. These examinees are classified as follows: 86examinees were from first level of ability (value of -3), 114examinees were from second level of ability (value of -2.5), 227examinees were from third level of ability (value of -2), 310examinees were from fourth level of ability (value of -1.5), 405examinees were from fifth level of ability (value of -1), 541examinees were from sixth level of ability (value of -0.5), 708examinees were from seventh level of ability (value of 0), 911examinees were from eighth level of ability (value of 0.5), 1183examinees were from ninth level of ability (value of 1), 1484examinees were from tenth level of ability (value of 1.5), 1693examinees were from eleventh level of ability (value of 2), 1884examinees were from twelfth level of ability (value of 2.5),and 2017examinees were from thirteenth level of ability (value of 3). Table 2 shows this information.

Table 2. Number of correct answers in each ability value

Ability Value	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
Number of correct answer	86	114	227	310	405	541	708	911	1183	1484	1693	1884	2017

Then, the proportion of correct answer,$P_O(\theta_j)$, was calculated for each ability level (Table 3).

Table 3. Proportion of correct answer in each ability value

Ability Value	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
Prop of correct answer	0.03	0.05	0.09	0.12	0.16	0.22	0.28	0.36	0.47	0.59	0.68	0.75	0.81

Now, the (θ_j, $P_O(\theta_j)$) pairs are plotted on the coordination screen (Fig.1).

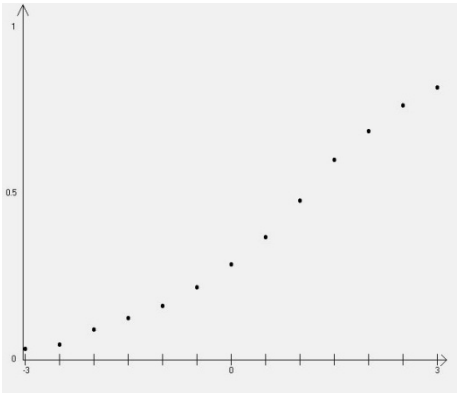

Figure 1. Plotted (θ_j, $P_O(\theta_j)$)pairs for item 16

Assuming b = 0.0, a = 1.0as priori values, the item characteristic curve will be drawn by$P(\theta) = \frac{1}{1+e^{1(0-\theta)}}$ function (Fig. 2).

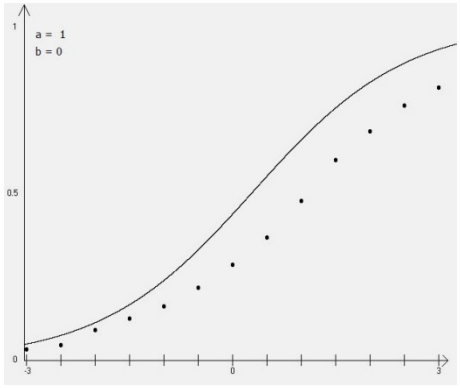

Figure 2. Item characteristic curve for item 16 by priori values

The sum of square of differences between $P_O(\theta_j)$ and $P(\theta_j)$ will be calculated:

$$\sum_{j=1}^{13}\left(P_o(\theta_j)-P(\theta_j)\right)^2 =$$

$$\left(\frac{86}{2500}-0.05\right)^2+\left(\frac{114}{2500}-0.08\right)^2+\cdots+\left(\frac{2017}{2500}-0.96\right)^2=0.36$$

The sum of square of difference for this curve is 0.36. By using the minimum sum of square of differences method, and examining all *b* and *a* values (from -3 to 3, and by 0.1 step), the b = 1.1and a = 0.8values, with sum of square of difference equal to 0.001, were accepted as the item 16 estimated parameters. The curve produced by these parameters is shown in Fig. 3.

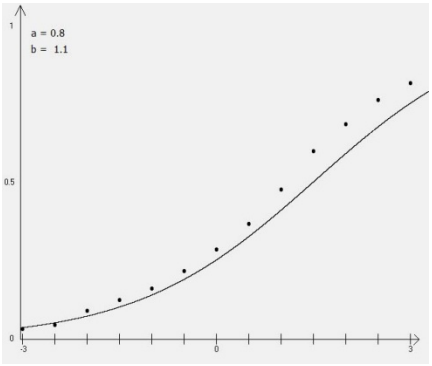

Figure 3. Item characteristic curve for item 16 by proper values

Finally the agreement of the observed proportions of correct response and those yielded by the fitted item characteristic curve for item 16 is measured by the chi-square goodness-of-fit index:

$$\chi^2=\sum_{j=1}^{13} m_j\frac{\left[P_o(\theta_j)-P(\theta_j)\right]^2}{P(\theta_j)Q(\theta_j)}$$

$$=2500\frac{\left[\frac{86}{2500}-0.04\right]^2}{0.04(1-0.04)}+2500\frac{\left[\frac{114}{2500}-0.05\right]^2}{0.05(1-0.05)}+\cdots$$

$$+2500\frac{\left[\frac{2017}{2500}-0.82\right]^2}{0.82(1-0.82)}=24.77$$

The standard chi-square value (by 12 degrees of freedom) with 0.01 confidence interval is equal to 26.22. Since the chi-square goodness-of-fit index for item 16 is less than the standard value, the estimated parameters will be accepted.

Results

Table 4 shows the estimated parameters of the items:

Table 4. The results of research (Number: Item's Number, *b*: Difficulty; *a*: Discrimination)

Number	*b*	*a*	Number	*b*	*a*
1	-2.7	0.3	31	2.1	1.5
2	-2.3	0.4	32	2.0	1.5
3	-2.4	0.2	33	-1.6	0.4
4	-2.0	0.4	34	-1.2	0.5
5	-1.5	0.5	35	1.4	0.9
6	-0.2	0.4	36	1.6	0.8
7	-0.5	0.3	37	1.5	0.8
8	-0.6	0.3	38	1.5	0.7
9	0.1	0.5	39	1.3	0.8
10	-2.0	0.3	40	1.1	0.7
11	-1.8	0.4	41	1.3	0.9
12	-1.9	0.3	42	1.0	0.8
13	-0.9	0.4	43	1.2	0.8
14	-0.9	0.4	44	1.1	0.8
15	-0.6	0.5	45	1.1	0.8
16	1.1	0.8	46	1.0	0.8
17	-0.1	0.5	47	1.9	1.4
18	1.5	0.9	48	2.6	1.8
19	-0.5	0.3	49	2.2	1.6
20	0.1	0.4	50	2.4	1.8
21	0.8	0.5	51	2.0	1.6
22	1.1	0.5	52	1.8	1.5
23	1.2	0.6	53	1.8	1.6
24	0.8	0.4	54	1.7	1.6
25	1.2	0.7	55	1.5	1.4
26	0.7	0.4	56	1.4	1.4

27	1.8	1.4	57	1.7	1.5
28	2.0	1.4	58	1.2	1.3
29	1.9	1.5	59	1.4	1.3
30	1.7	1.3	60	1.2	1.1
			61	1.0	1.1

Conclusion

The first step for designing an adaptive test is developing an item bank. An item bank with at least 60 items can reasonably support a two-parameter test (Hortensius & Weiss, 2012).

In this paper an effort has been made to develop an item bank for homogeneous second order differential equations. To accomplish this goal, 61 questions were created and calibrated through a simulated test. After estimating the item's parameters, their accuracy has been tested by the chi-square goodness-of-fit index.

For refining this research's results, one can use these items in a real-world test, and after comparing the results, make the appropriate changes in the parameters' values.

Developing calibrated item banks is crucial for designing adaptive tests. Thus, adding or refining the items introduced in this research is greatly appreciated.

References:

Babcock, B., and Weiss, D. J. Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In *D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 2009.

Baker, F. B. The Basics of Item Response Theory. *ERIC Clearinghouse on Assessment and Evaluation*,2001.

Hortensius, L., and Weiss D. J. Small item bank CAT, Interaction of scoring method, termination criteria, and item bank shape. *International Association for Computerized Adaptive Testing Conference, Sydney, Australia*, 2012.

Sympson, J. B. Evaluating the Results of Computerized Adaptive Testing. University of Minnesota, 1970.

Thompson, N. A., and Weiss D.J. A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation.* 16(1),2011.

van der Linden, W. J. A Comparison of Item-Selection Methods for Adaptive Tests With Content Constraints. *Law School Admission Council Computerized Testing Report 04-02*, 2005

Weiss, D. J. Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. *Measurement and Evaluation in Counseling and Development*. 37(2),2004.
Weiss, D. J. Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences*. 2(1), 2011.